International Journal of Research in Marketing 36 (2019) 492–508

Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media

Susan A.M. Vermeer [a,*,1], Theo Araujo [a,1], Stefan F. Bernritter [b,c,1], Guda van Noort [a,1]

[a] University of Amsterdam, Amsterdam School of Communication Research (ASCoR), P.O. Box 15793, 1001 NG Amsterdam, the Netherlands
[b] King's Business School, King's College London, Bush House, 30 Aldwych, London WC2B 4BG, United Kingdom
[c] Goldsmiths, University of London, Institute of Management Studies, New Cross, London SE14 6NW, UK

ARTICLE INFO

Article history:
First received on March 5, 2018 and was under review for 3 months
Available online 11 February 2019

Keywords:
eWOM
Webcare
Social media
Digital marketing strategies
Automated content analysis
Sentiment analysis
Machine learning

ABSTRACT

The increasing volume of firm-related conversations on social media has made it considerably more difficult for marketers to track and analyse electronic word-of-mouth (eWOM) about brands, products or services. Firms often use sentiment analysis to identify relevant eWOM that requires a response to consequently engage in webcare. In this paper, we show that sentiment analysis of any kind might not be ideal for this purpose, because it relies on the questionable assumption that only negative eWOM is response-worthy and it is not able to infer meaning from text. We propose and test an approach based on supervised machine learning that first decides whether eWOM is relevant for the brand to respond, and then—based on a categorization of seven different types of eWOM (e.g., question, complaint)—classifies three customer satisfaction dimensions. Using a dataset of approximately 60,000 Facebook comments and 11,000 tweets about 16 different brands in eight different industries, we test and compare the efficacy of various sentiment analysis, dictionary-based and machine learning techniques to detect relevant eWOM. In doing so, this study identifies response-worthy eWOM based on the content instead of its expressed sentiment. The results indicate that these machine learning techniques achieve considerably higher accuracy in detecting relevant eWOM on social media compared to any kind of sentiment analysis. Moreover, it is shown that industry-specific classifiers can further improve this process and that algorithms are applicable across different social networks.

1. Introduction

"@Ben&Jerry's, I would love to have some kind of coffee ice cream again. So tasty, pity that these are always being removed from the assortment. When will a new flavour be introduced?" This is a quote taken from the social media data used in this study. Should Ben & Jerry's respond to this consumer's social media post? Common sense might suggest that this message is important to answer. This notion finds support in research that shows that it is beneficial for brands to respond to positive messages like this (e.g., Schamari & Schaefers, 2015). Yet, most automated social media monitoring approaches would not be able to identify this post as being relevant, because these approaches rely on sentiment extraction (i.e., a technique that aims to determine the extent

* Corresponding author.
E-mail addresses: S.A.M.Vermeer@uva.nl, (S.A.M. Vermeer), T.B.Araujo@uva.nl, (T. Araujo), stefan.bernritter@kcl.ac.uk, (S.F. Bernritter), G.vanNoort@uva.nl. (G. van Noort).
[1] Note: All authors contributed equally to this article.

https://doi.org/10.1016/j.ijresmar.2019.01.010

to which a text is negative, positive or neutral; cf., Pang & Lee, 2008) and usually prioritize negative eWOM. Moreover, a high percentage of eWOM is irrelevant for the firm and can be categorized as clutter. For instance, consumers often tag friends, write unrelated comments, or post GIF's in response to firms' posts. Approaches based on sentiment extraction would still categorize this type of eWOM as either negative, positive, or neutral, while the content does not merit or require a brand response. This eventually will result in both an inefficient categorization of eWOM and unnecessary manual labour for the webcare staff filtering out irrelevant content. As a result, many marketers experience difficulties with automatically identifying relevant and response-worthy eWOM (Grégoire, Salle, & Tripp, 2015).

In the present study, we argue that because of the described shortcomings, sentiment extraction is conceptually inexpedient as a means to discover relevant eWOM in social media. While there is a huge need in the industry for more reliable tools to indicate eWOM that should be addressed by the brand (Humphreys & Wang, 2017), this conceptual inappropriateness of sentiment analysis for webcare purposes has not been addressed in the literature so far. As an alternative to sentiment extraction, the central aim of this study is to explore how supervised machine learning methods that focus on context and relevance of content instead of its sentiment can optimize firms' efficacy in automatically identifying relevant eWOM on social media, as compared to any type of sentiment analysis and a dictionary-based approach. Additionally, we investigate the roles of platform type and industry in this context.

This paper contributes to the literature and managerial practice in several ways. First, we demonstrate that even advanced types of sentiment analysis are not accurate enough in recognizing eWOM content that is *in need of a response* by the brand. Our machine learning-based categorization showed that focussing on eWOM type before calculating satisfaction yields considerably more accurate results than any type of sentiment analysis. Secondly, we extend the generalizability of our findings by demonstrating that our approach—which has been developed based on Facebook data—outperforms any type of sentiment analysis also conducted on Twitter data. This shows that our approach is dynamic enough to be quickly adopted in new venues, which has been demonstrated to be highly problematic in previous research (Schweidel & Moe, 2014). Thirdly, we provide novel insights into the suitability of different approaches of training algorithms for eWOM monitoring by comparing the effectiveness of 11 text classification algorithms (i.e., four types of sentiment analysis, a dictionary-based approach, and six machine learning algorithms) on detecting relevant eWOM for 16 brands. Finally, we explore under which circumstances machine learning models need to be trained with data specific to a given industry and under which circumstances a generic classifier may be sufficient, and therefore more cost-effective.

2. Conceptual framework

2.1. The importance of identifying and responding to relevant eWOM

The "act of engaging in online interactions with (complaining) consumers, by actively searching the web to address consumer feedback (e.g., questions, concerns, and complaints)" is referred to as *webcare* and has become a central part of firms' customer relationship management strategies (Van Noort & Willemsen, 2012, p. 133). By engaging in webcare, brands show that they care for their consumers (Bhandari & Rodgers, 2018). This may benefit brands in two ways. First, it can counteract possible negative outcomes of negative eWOM. As such, webcare has been demonstrated to prevent negative eWOM from backfiring or evolving into crises (Van Noort, Willemsen, Kerkhof, & Verhoeven, 2014) and to reduce failure attributions of complaining customers (Weitzl, Hutzinger, & Einwiller, 2018). Previous studies have also shown that webcare is effective in positively influencing potential consumers who are exposed to negative comments posted by other customers (Willemsen, Neijens, & Bronner, 2013). Second, webcare can be used as an effective marketing tool by engaging with positive eWOM. Schamari and Schaefers (2015), for instance, showed that that webcare can increase consumers' positive engagement on consumer-generated platforms. Supporting this notion, Colliander, Dahlén, and Modig (2015) demonstrated that engaging in dialogue with consumers on social media increases consumers' brand evaluation, because consumers perceive brands that engage in dialogic communication to be caring. Accordingly, in order to restore, extend, and maintain a brand's reputation and its relationships with customers (Coombs, 2002), it is essential to track and act upon relevant eWOM. We define eWOM to be relevant in a webcare context if it includes an expression that concerns the product, the service and/or the entire firm (cf., Huibers & Verhoeven, 2014).

2.2. Social media monitoring: going beyond valence

eWOM can be either negative, positive, or neutral. Social media monitoring in a customer relationship management context is often based on this distinction, with the aim of filtering out negative eWOM in order to handle customer complaints (e.g., Van Laer & De Ruyter, 2010; Van Noort & Willemsen, 2012). This approach assumes that firms want to avoid negative eWOM as much as possible, so it should be approached with more urgency than other types of eWOM. Indeed, the spread of negative eWOM can cause costly or irreparable damage for brands (Kietzmann & Canhoto, 2013) as it has the ability to influence all stages of the consumer decision-making process as well as brand perception and evaluation (Van Noort & Willemsen, 2012). Furthermore, studies examining webcare report various examples of brands that have suffered massive reputation loss as a result of negative eWOM (e.g., Van Laer & De Ruyter, 2010). Thus, if unresolved, online complaints voiced through negative eWOM can potentially have detrimental consequences for brands' reputation.

However, there has been an ongoing discussion for at least the last decade about the extent to which valence of eWOM affects brand outcomes. Results are generally mixed, but the findings of a recent meta-analysis question the assumption that negative

eWOM is inherently harmful to brands (Babić Rosario, Sotgiu, De Valck, & Bijmolt, 2016). These authors showed that negative-valenced eWOM only has a negative effect on sales in the later stages of the product life cycle and for low-financial-risk products. Moreover, in a recent study, Wilson, Giebelhausen, and Brady (2017) demonstrated that negative eWOM can increase behavioural intentions for consumers with a high self-brand connection. In the same vein, Berger, Sorensen, and Rasmussen (2010) showed that negative eWOM can actually increase sales of rather unknown products because it increases product awareness. Negative eWOM is thus not necessarily harmful to brands and might under certain circumstances even be beneficial.

Another disadvantage of a social media monitoring approach that merely focusses on negative eWOM is that it neglects the large number of consumers that reach out to or talk about brands in a positive way. This positive type of eWOM has been demonstrated to have positive effects on various consumer mindset metrics and sales (e.g., Pauwels, Aksehirli, & Lackman, 2016), especially on earned social media (Colicev, Malshe, Pauwels, & O'Connor, 2018). Engaging with these consumers can greatly benefit brands which thus constitutes an effective social media marketing tool (Schamari & Schaefers, 2015). But also findings concerning the consequences of positive valence in eWOM are not equivocal. Recent studies in the domain of customer reviews, for example, showed that too positive reviews can actually result in a decrease in sales (Maslowska, Malthouse, & Bernritter, 2017) and behavioural intentions (Kupor & Tormala, 2018). Furthermore, neutral eWOM has also been found to play an important role in affecting consumer behaviour. This is because it can change how consumers perceive positive and negative eWOM (Tang, Fang, & Wang, 2014). Thus, it seems reasonable to conclude that positive, neutral, and negative eWOM can all be relevant for brands and should therefore be monitored.

This illustrates that any type of sentiment analysis, by definition, cannot be the optimal solution to track relevant eWOM for responses by the brand. If all three types of sentiment are potentially relevant, how can sentiment extraction help in identifying what is relevant after all? In the current paper, we therefore propose an approach that goes beyond valence of eWOM, and rather uses supervised machine learning to automatically identify eWOM in which action by the brand may be necessary.

2.3. Previous research on automated content analysis of eWOM

Many marketers experience difficulties with identifying relevant and response-worthy eWOM (Grégoire et al., 2015). This might be a result of the conceptual inexpedience of sentiment extraction for identifying relevant eWOM. Ongoing advances in computational methods and natural language processing provide opportunities for solving the managerial challenge of processing large-scale, and potentially nearly real-time streams of eWOM. We present here a short overview of the main approaches.

2.3.1. Unsupervised machine learning

On the one hand, when it comes to identifying the topics that emerge out of the potentially millions of messages that are generated by consumers about brands, unsupervised machine learning methods such as latent Dirichlet allocation (LDA) topic modelling have shown great potential. In this way, researchers interested in describing frames or topics, without having any predefined categories, can use unsupervised machine learning to make sense of unstructured data. For example, Tirunillai and Tellis (2014) have demonstrated how LDA topic models are able to identify key latent dimensions of consumer posts at aggregated levels, and in particular analyse aspects related to quality within and across brands. Büschken and Allenby (2016) have shown that the same technique is also able to identify topics that emerge out of a sample of consumer reviews. LDA topic models, and other unsupervised machine learning techniques, therefore, present potentially optimal solutions for marketing managers that may want to know, in *aggregate levels*, what themes or topics consumers discuss about brands (e.g., dimensions of quality as in Tirunillai & Tellis, 2014; "real pizza", menu, return, food ordered, service and staff, as in Büschken & Allenby, 2016). These techniques, however, say little about which specific content *requires a response*. For example, several posts may be aggregated by an LDA topic model within the same topic (e.g., service and staff), yet only a small share of these posts may actually be complaints, questions, suggestions (among other eWOM types), which would actually require a response by the brand.

On the other hand, supervised machine learning and other classification techniques (e.g., sentiment analysis, dictionary-based approaches) offer potential solutions for marketing managers interested in categorizing specific messages in pre-defined categories (such as negative, positive or neutral or, in our case, response-worthy eWOM) so that action can be taken towards specific messages. We discuss these techniques below.

2.3.2. Sentiment analysis

Sentiment analysis has the advantage of providing a relatively easy to interpret, off-the-shelf metric of valence of eWOM (e.g., the extent to which the consumer post is composed of positive-, neutral-, or negative-valenced words) however, it has important limitations if applied in a webcare context. First, because eWOM messages are characterized by unstructured text formats, informal speak and simplified expressions, social media monitoring tools that are based on online mentions or sentiment analyses might miss important gaps when analysing eWOM content (Pai, Chu, Wang, & Chen, 2013). As such, retrieving information from consumer opinions is typically a challenging task. As many social media monitoring tools rely on sentiment, a set of disruptive factors pose challenges to the accuracy and the usefulness of this technique for eWOM classification (Pang & Lee, 2008). Determining the sentiment of user-generated content (UGC) is particularly difficult in instances in which the literal meaning of the text is not the intended meaning of the content (e.g., Kunneman, Liebrecht, Van Mulken, & Van Den Bosch, 2015). Furthermore, personal pronouns (e.g., *you*, *she*, *it*, etc.), adverbs of negation (e.g., *neither*, *never*, *none*, etc.) and adjectives of quantity (e.g., *many*, *enough*, *little*, etc.) can negatively affect the accuracy of sentiment analysis (e.g., Munoz-Garcıa & Navarro, 2012). Finally, the efficiency of sentiment analysis is likely to differ across brands and industries, as people use varied language to describe experiences in different

domains (Owsley, Sood, & Hammond, 2006). Given these limitations, brands call for an appropriate method to effectively identify relevant eWOM, which might be based on the computational treatment of sentiment as well as opinion and subjectivity in the text (Humphreys & Wang, 2017).

Another limitation, we argue, is that when responding to eWOM it is important to consider the content of the eWOM post. The content of eWOM has been shown to have important effects on its efficacy. For example, content factors such as explicitness of endorsements in online reviews (Packard & Berger, 2017), argument diversity and density (Willemsen, Neijens, Bronner, & De Ridder, 2011), sentiment diversion (Zhang, Li, & Chen, 2012), review subjectivity (Ghose & Ipeirotis, 2011), and valence orientation (Gopinath, Thomas, & Krishnamurthi, 2014) have all important consequences for how eWOM is perceived by others and how it influences them. Sentiment analysis is not able to identify the eWOM content and will therefore suffer from inaccurate identification of what is relevant to respond to and what is not. This is echoed by previous research that already highlighted the importance of tracking eWOM content, as this generates more important insights than quantity (i.e., volume) or sentiment (Godes & Mayzlin, 2004; Pauwels et al., 2016). Especially when it comes to social media monitoring in a webcare context, focusing on content instead of valence or volume might be thus be a more worthwhile approach.

2.3.3. Dictionary-based approach

Dictionary-based text analysis can be applied to detect the presence of certain words in order to arrive at a webcare prediction. The basic idea of a dictionary-based approach is that researchers manually assign lists of keywords that correspond to groupings (e.g., topics, attributes, etc.) that they hope to identify in the text (see e.g., Pennebaker, Boyd, Jordan, & Blackburn, 2015). Each unit of analysis (e.g., paragraph, sentence; in our case eWOM content) is scanned for the presence of those words. If a match is found, then the unit is annotated as containing that grouping.

Compared to manual content analysis, a dictionary-based approach increases efficiency of text classification tasks to a great extent (Guo, Vargo, Pan, Ding, & Ishwar, 2016). Researchers may use preset word lists, as these are often generic across domains and can be extended by custom word lists (Hartmann, Huppertz, Schamp, & Heitmann, 2018). While sentiment analysis tools often rely on dictionary-based approaches to determine the sentiment of posts based on lists of words with negative or positive valence (e.g., LIWC, SentiStrength), dictionary-based approaches can be used to create dictionaries with words specific to the topic of interest (e.g., detection of webcare posts), thus potentially being able to provide better performance than sentiment analysis for the task at hand. A good number of recent studies have employed this method to analyse social media data for understanding social media firestorms (see e.g., Hansen, Kupfer, & Hennig-Thurau, 2018), and product reviews (see e.g., Moon & Kamakura, 2017). However, just like sentiment analysis, it is difficult for dictionaries to infer meaning from co-occurrences of words. Additionally, as the process relies on various subjective steps, dictionary-based text analysis often risks being over-specific and missing words; hence, it might not adequately reflect the entire data set, or be as flexible to be applied to new texts.

2.3.4. Supervised machine learning

Machine learning techniques can handle more complex meaning compared to dictionary-based approaches. Unlike pre-trained sentiment analysis algorithms that look for the manifest valence of a text, a supervised machine learning algorithm learns from a human coder's decisions and would allow marketers to solve the classification problem for an unlimited amount of eWOM messages. While requiring manual labour initially, as training a supervised machine learning algorithm requires an existing dataset of eWOM texts and their classification (response-worthy, or not), this approach can be highly useful for coding implicit variables in a large dataset. Moreover, as it does not start with pre-existing assumptions (e.g., that eWOM must contain negative words, as mostly is the case with sentiment analysis), this approach is generally able to handle complex meaning, and allows a large amount of flexibility for the firm in terms of which categories can be classified. As such, using supervised machine learning does not only increase efficiency, but also transparency and reproducibility (Boumans & Trilling, 2016). Further, classifiers can be trained so they can be used over and over again.

Many different approaches exist for classification tasks,[2] varying from very simple algorithms to more computationally demanding classifiers. *Bayesian algorithms* (e.g., Bernoulli Naïve Bayes, Multinomial Naïve Bayes), for example, are simple, probabilistic algorithms that are often used for text classification (e.g., Dhillon, Mallela, & Kumar, 2003; Tirunillai & Tellis, 2012) that still tend to perform well. Whereas such classifiers are commonly known for speed, efficiency and computational power savings (see Kübler, Wieringa, & Pauwels, 2017), more advanced algorithms, often yield better results. *Support Vector Machines* (SVMs), for example, often outperform Bayesian algorithms as they are large-margin rather than probabilistic. SVMs use a subset of training points in the decision function (i.e., support vectors), which makes them memory efficient, and allows them to avoid overfitting (Cortes & Vapnik, 1995). SVMs are often used in marketing research, as they are highly effective at traditional text classification (see e.g., Kübler, Colicev, & Pauwels, 2017; Li & Wu, 2010; Tirunillai & Tellis, 2012). Furthermore, the Passive Aggressive (PA) algorithm works similarly to SVM. It can be viewed as the online version of an SVM, as the PA algorithm uses the margin to update the classifier. In other words, if the prediction is correct it will be passive, whereas the weights will be updated for correct classification when the prediction is wrong. In this way, the PA algorithm is able to reach high accuracy in text classification, such as online product reviews (see Cui, Mittal, & Datar, 2006). Finally, the *Stochastic Gradient Descent* (SGD) has been successfully applied to large-scale corpora with data sparsity, which regularly tend to present challenges for machine learning especially when it

[2] We have also tested the usage of a dictionary-based approach, with a dictionary created out of webcare-related words (as outlined below). We thank the anonymous reviewers for this valuable suggestion.

comes to natural language processing and text classification (see e.g., Martínez-Cámara, Martín-Valdivia, Urena-López, & Montejo-Ráez, 2014).

Supervised machine learning is increasingly applied in marketing and consumer research. For instance, Okazaki, Diaz-Martin, Rozano, and Menendez-Benito (2015) used a supervised machine learning approach to classify customers' emotional state and dialogue acts in UGC on Twitter. Furthermore, in a recent study, Ordenes et al. (2018) used SVMs to mine brands' message intentions on social media and consequently assessed their effects on consumers' sharing behaviour. Supervised machine learning is also often used to analyse the sentiment of UGC. Homburg, Ehm, and Artz (2015), for example, used supervised SVM learning to investigate the effects of firm interventions in online forums on consumer sentiment. Tirunillai and Tellis (2012) used a semi-parametric SVM to determine the valence of customers' online reviews. Moreover, in a recent study Kübler, Colicev, and Pauwels (2017) compared the dynamic explanatory power of different SVM to that of dictionary-based analysis of sentiment-rich words and volume measures. A systematic overview of related research can be found in Web Appendix A.

2.4. The current study

Relying on supervised machine learning, we demonstrate how such models can help filter clutter out of eWOM messages before proceeding with classifications that allow us to predict the underlying meaning of content and therefore determine the type of eWOM. In particular, we create and test the efficacy of supervised machine learning models that first indicate whether eWOM is relevant for the firm, and—based on seven different types of eWOM (i.e., rejection, complaint, comment, question, suggestion, acknowledgement and compliment; cf., Brown & Levinson, 1987)—classify three dimensions of customer satisfaction. The latter can help managers to still prioritize eWOM with a certain level of customer satisfaction and allows us to compare our approach with more traditional approaches of sentiment extraction. The key difference here is that our aggregation of customer satisfaction is not based on whether eWOM contains words that are thought to have a certain valence, or on whether an (machine learning) algorithm decided that the text is indeed positive/neutral/negative, but that we infer satisfaction from the meaning of the post instead of using sentiment as a proxy. Importantly, this approach allows us to filter out clutter before engaging in eWOM categorization, which should increase the accuracy and efficiency of the algorithms. A schematic outline of our process for identification of response-worthy eWOM can be found in Fig. 1.

3. Method

3.1. Sample

A multinational media and digital marketing communications organization provided actual consumer comments on Facebook. Additionally, we collected tweets via Coosto, 2017 (see Appendix D for detailed information about the data collection for Twitter). We used the Global Industry Classification Standard to select sixteen brands across eight different industries (i.e., two brands per industry), namely (1) Automobiles & Components (e.g., manufacturers of parts and accessories for automobiles and motorcycles), (2) Consumer Durables & Apparel (e.g., manufacturers of electric household appliances and related products), (3) Consumer Services (e.g., operators of casinos and gaming facilities), (4) Food, Beverage & Tobacco (e.g., producers of alcohol or non-alcoholic beverages), (5) Household & Personal Products (e.g., producers of non-durable household products), (6) Insurance (e.g., insurance and

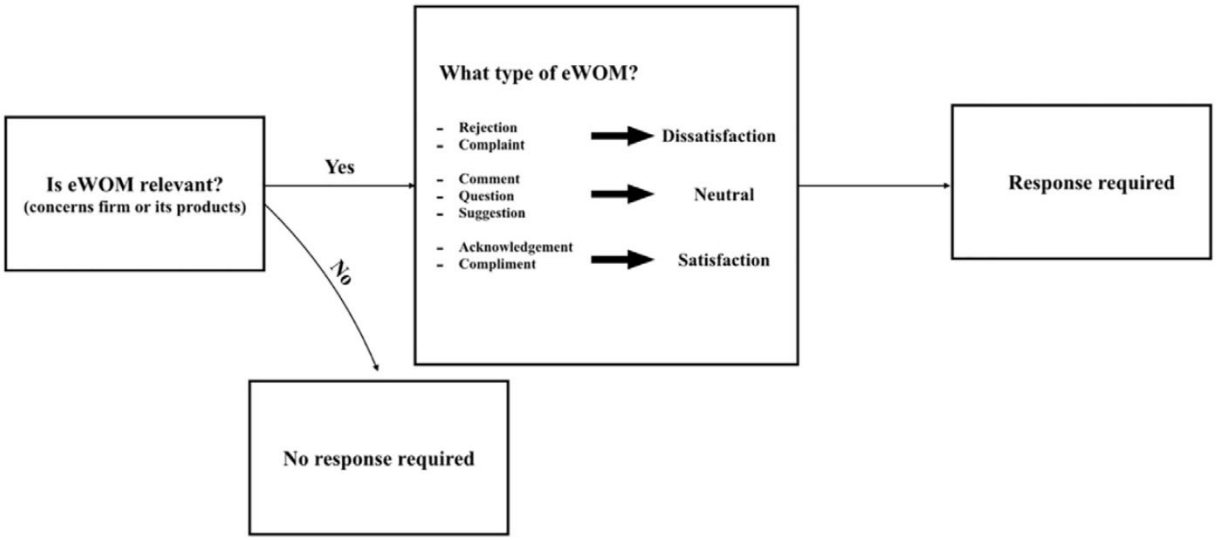

Fig. 1. Proposed process for response-worthy eWOM identification.

reinsurance companies), (7) Retailing (e.g., operators of stores offering diversified general merchandise) and (8) Telecommunication Services (e.g., operators providing wireless and fixed-line telecommunications services; Phillips & Ormsby, 2016).

3.2. Coding and procedure

Based on this data set, we developed several models to examine the effectiveness of machine learning techniques. To do so, we randomly sampled 5% of the comments made on the Facebook brand pages ($N = 60{,}150$) and Twitter ($N = 11{,}154$). This sample was categorized by three human coders. Firstly, the coders had to evaluate the relevance of every message (i.e., "The eWOM content includes an expression that concerns the product, the service and/or the entire brand. Hence, a webcare response of the brand is necessary or appropriate", Huibers & Verhoeven, 2014). Then, every relevant message was coded for the presence of one or multiple types of eWOM, specifically: (1) Rejection (i.e., the consumer discards the product, the service and/or the entire brand), (2) Complaint (i.e., the consumer complaints or expresses criticism towards the product, the service and/or the entire brand), (3) Comment (i.e., the consumer expresses their thoughts regarding the product, the service and/or the entire brand), (4) Question (i.e., the consumer poses a question about the product, the service and/or the entire brand), (5) Suggestion (i.e., the consumer suggests an idea regarding the product, the service and/or the entire brand), (6) Acknowledgement (i.e., the consumer expresses gratitude and/or appreciation regarding the product, the service and/or the entire brand) and (7) Compliment (i.e., the consumer expresses appreciation towards the product, the service and/or the entire brand; Brown & Levinson, 1987). As each message could be categorized into more than one eWOM type, the coders also indicated which type of eWOM was most dominantly present.

Ultimately, as identifying satisfied and dissatisfied consumers can help us to understand the content and potentially the urgency of the comments (Okazaki et al., 2015), and as it helps us to compare our approach to sentiment analysis that is based on a negative/positive/neutral distinction, we aggregated the types of eWOM in three main categories: Dissatisfaction, Satisfaction, and Neutral. Firstly, the category Dissatisfaction, consists of consumers expressing a rejection and/or a complaint regarding the product, the service and/or the entire brand. Secondly, the Satisfaction category includes acknowledgements and/or compliments. Finally, anything between the two extremes, including comments, questions and/or suggestions, was classified as Neutral.

3.2.1. Intercoder reliability Facebook

In the first phase of the coding process, two coders categorized 3900 Facebook comments to assess intercoder reliability. Cohen's κ was sufficient for all categories, as the minimal threshold of 0.70 was reached. The coders discussed their interpretations of conflicting results and a consensus was reached. On this basis, the coders categorized the additional 56,250 Facebook comments (see Web Appendix B, C and D for detailed information about the coding procedure).

3.2.2. Intercoder reliability Twitter

Furthermore, two coders categorized 1200 tweets to assess intercoder reliability. Cohen's κ was sufficient for almost all categories. The coders discussed their interpretations of conflicting results and a consensus was reached. On this basis, one coder categorized the remaining tweets.

3.3. Variables

We present the results of the manual coding in Table 1. We use the manually coded Facebook data to build and test our classifiers. To increase the generalizability of our results, we use the Twitter data to test the classifiers and then examine the differences across social media platforms.

Table 1
Descriptive statistics: relevance and eWOM type based on manual coding.

Variable	Facebook (in %)	Twitter (in %)
Relevance		
Irrelevant eWOM	85.7	71.1
Relevant eWOM	14.3	28.9
eWOM type		
Dissatisfied		
Rejection	2.9	5.4
Complaint	12.9	35.8
Neutral		
Comment	27.1	18.3
Question	11.8	25.2
Suggestion	1.3	1.7
Satisfied		
Acknowledgement	1.9	2.9
Compliment	42.1	10.6

Note. Table 1 presents the dominant eWOM Type; hence this adds up to 100%. Unfortunately, merely 91 cases of sarcasm were found in the training set, and we decided to exclude this variable for further analyses.

3.3.1. Dependent variables

First, *relevance* is the dependent variable when exploring machine learning models to detect relevant eWOM. As shown in Table 1, 14% of Facebook comments and 29% of tweets include an expression that concerns the product, the service and/or the entire brand and a webcare response of the brand is necessary or appropriate. The second dependent variable of this study is *eWOM type*, to examine the extent to which models using machine learning can detect urgent relevant eWOM more effectively compared to a generic classifier. As indicated in Table 1, the majority of the Facebook comments represent satisfied consumers, namely 44%, whereas for Twitter the majority concerns dissatisfied consumers. We treat messages indicating dissatisfaction (i.e., rejection, complaint) as urgent cases to be able to use it as a benchmark against sentiment analyses.

3.3.2. Independent variable

The independent variable of this study was the actual *content* (i.e., the text) of the message, which was used as the basis for training the machine learning classifiers. Before training each classifier, the text was converted to a bag-of-words model and was used as the input for the model. Different pre-processing steps have been used resulting in three different text categories. The first category includes the original content of the eWOM message, for example:

> "*Thanks [brand], I am so delighted with this product!*".

Next, Dutch stop words such as articles (e.g., *the, a* and *an*), personal pronouns (e.g., *I, me* and *he*), coordinating conjunctions (e.g., *for, but* and *so*) and prepositions (e.g., *in, towards* and *before*) were removed.[3] The list of stop words was based on using the Python Natural Language Tool Kit (NLTK) package (Loper & Bird, 2002). The stop word removal means that the text category results in:

> "*Thanks [brand], am delighted product!*".

Finally, explicit mentions of the brand have been removed from the eWOM message, aiming at creating classifiers that could be brand-independent. This results in the third text category:

> "*Thanks brandname000, am delighted product!*".

3.4. Algorithms to detect eWOM messages

Based on this dataset, models have been generated with various automated content analysis techniques. These include sentiment analysis, dictionary-based identification of eWOM categories and machine learning techniques.

3.4.1. Sentiment analysis

We test the performance of various sentiment analysis techniques, namely: (1) Linguistic Inquiry and Word Count (LIWC), (2) Pattern, and (3) SentiStrength (operationalized in two ways). Web Appendix B provides more details about these different techniques.

3.4.2. Dictionary-based approach

We also incorporated a dictionary-based approach, to detect the presence of certain words in order to arrive at a webcare prediction. We combined two different approaches: (1) an inductive (i.e., tf-idf scores) and (2) a deductive approach (i.e., from our codebook). Based on the Facebook comments, we make use of tf-idf (i.e., an algorithm frequently applied in information retrieval and text mining, which measures how common a word is across an entire collection of texts; see e.g., Zhang, Yoshida, & Tang, 2011). Based on the tf-idf scores, we selected words that are most likely to identify relevant eWOM. That is, we retrieved the 75 words with the highest tf-idf per each of the seven eWOM categories. From these 75 words, we selected those that are most likely to identify relevant and urgent eWOM in need of a webcare response. Consequently, we added words (and synonyms) from our codebook to the list of words. Combining the results from the inductive and deductive approach, resulted in a list of 150 words. Web Appendix C provides more detailed information about this technique.

3.4.3. Machine learning

Based on our Facebook data, we then created models using six machine learning algorithms to compare the performance of machine learning with the sentiment analysis output. When a classifier is trained and tested in this manner, we refer to it as a generic classifier. We selected the following classifiers: (1) Multinomial (MNB) and (2) Bernoulli (BNB) classifiers, (3) Logistic Regression (LR), the (4) Stochastic Gradient Descent (SGD), (5) Support Vector Machines (SVM) and (6) the Passive Aggressive (PA) algorithm. Overall, MNB, BNB, LR, SGD, SVM and PA classifiers have been widely used in prior research on social media data mining (Kaiser & Bodendorf, 2012). For each machine learning algorithm, we created three models based on how the text could be processed: (1) the original text of the comment, (2) the text with the stop words removed, and (3) the text with the stop

[3] Translations of the actual Dutch words are presented here as examples for the ease of the reader.

words and the brand name removed. Web Appendix D discusses in detail the different approaches of machine learning algorithms and the training data used.

3.4.4. Classification evaluation of algorithms

The algorithms and sentiment analysis have been compared based on their *precision, recall* and *accuracy* to determine their performance. These measures are based on the following concepts associated with how an occurrence has been correctly or incorrectly classified (Esuli & Sebastiani, 2010):

- True positive: the occurrence has been correctly classified as part of the category;
- False positive: the occurrence has been incorrectly classified as part of the category;
- True negative: the occurrence has been correctly classified as not part of the category;
- False negative: the occurrence has been incorrectly classified as not part of the category.

The measures of accuracy, precision and recall are defined as follows: The accuracy (i.e., F-measure) is a weighted arithmetic metric that considers precision as well as recall (Esuli & Sebastiani, 2010; Kent, Berry, Luehrs Jr., & Perry, 1955).

$$Accuracy = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

Precision (also called positive predictive value) is used to measure when an occurrence that belongs to the category set is classified as part of the category set (Esuli & Sebastiani, 2010; Kent et al., 1955).

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{2}$$

Recall (also called sensitivity) measures when an occurrence is rightly classified according to its category (Esuli & Sebastiani, 2010; Kent et al., 1955).

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{3}$$

4. Results

The results of machine learning techniques are often difficult to interpret into managerial insights (Nakhaeizadeh, Taylor, & Kunisch, 1997). In order to take advantage of the features that machine learning techniques have to offer, and more importantly to decide which of the algorithms has performed better, we also take managerial considerations into account when interpreting the results. As mentioned before, precision indicates how correct the results of the classifier are in terms of predicting whether an item belongs to the category, while recall reveals their completeness (i.e., to what extent all members of a category could be detected by the classifier). In other words, precision implies the proportion of detected relevant eWOM that are actually relevant eWOM. Thus, when a brand needs accurate results, for example to save time and/or money, a higher precision is necessary. Recall, on the other hand, indicates what proportion of relevant eWOM were classified by us as relevant eWOM. Thus, when it is crucial for a brand to identify a majority of all relevant eWOM, a higher recall is desirable.

4.1. Detecting relevant eWOM

First, we examined how sentiment analysis, a dictionary-based approach, and machine learning perform when trying to detect relevant eWOM. The results of the text classification categories indicate that removing stop words as well as the brand name from the eWOM message produced the best accuracy (0.50), followed by the original text (0.48) and merely stop word removal (0.45). Further, the accuracy, precision and recall for all methods have been calculated for a comparative purpose. The results presented are merely part of the test set. Table 2 summarizes the results of a generic classifier via the different methods.

4.1.1. Accuracy

Firstly, the results for the sentiment analysis show that the accuracy for effectively detecting relevant eWOM are between 0.15 and 0.28. Subsequently, the results gradually improve when employing the machine learning techniques. The accuracy of the various algorithms ranged from 0.46 to 0.60. The best classification method was LR[4] (0.60), followed by SVM (0.59) and PA (0.56).

[4] By using the Akaike Information Criterion (AIC; the relative quality of a model for a given set of data), we estimated the quality of each algorithm relative to each of the other algorithms. By doing so, we found decisive evidence in favour of the LR model relative to the other models ($AICc = 7217.03$; see Web Appendix F; Table F1 for more details).

Table 2
Model generation results of the generic classifier (Relevance).

Technique ($N = 12{,}030$)	Accuracy	Precision	Recall
Sentiment analysis			
LIWC	0.15	0.35	0.10
P	0.17	0.26	0.13
S	0.28	0.33	0.25
SN	0.25	0.34	0.20
Dictionary-based			
D	0.24	0.33	0.18
Machine learning			
BNB	0.46	0.49	0.43
MNB	0.47	0.61	0.38
LR	0.60	0.48	0.79
SGD	0.54	0.47	0.78
SVM	0.59	0.47	0.78
PA	0.56	0.47	0.69

Note. LIWC *Linguistic Inquiry and Word Count;* P *Pattern;* S *Sentiment Negative*; SN *Sentiment Net*; D *Dictionary-based;* BNB *Bernoulli Naïve Bayes*; MNB *Multinomial Naïve Bayes*; LR *Logistic Regression*; SGD *Stochastic Gradient Descent*; SVM *Support Vector Machine*; and PA *Passive Aggressive*. Performance scores ≥0.60 have been highlighted. Results merely derived from the test set.

4.1.2. Precision

The precision of the various machine learning techniques varies between 0.47 and 0.61. The MNB indicates the best performance (0.61), followed by BNB (0.49) and LR (0.48).

4.1.3. Recall

Furthermore, the recall of the various machine learning techniques varies between 0.38 and 0.79. The best classification method was LR (0.79), followed by SGD (0.78) and SVM (0.78). Overall, the MNB performs better on precision (0.61), whereas the LR performs better on recall (0.79). Thus, when it is crucial for a brand to identify all relevant eWOM, the LR can be argued as the most appropriate.

4.2. Detecting type of eWOM

Second, we examined whether models using machine learning can detect urgent relevant eWOM more effectively as compared to sentiment analysis or a dictionary-based approach. The results of the text classification categories indicate that on average the original text produced the best classification results (0.34), followed by both stop word and brand name removal (0.33) and merely stop word removal (0.32). The accuracy, precision and recall of a classification per eWOM type are presented in Table 3.

4.2.1. Accuracy

Firstly, the results for the sentiment-net classifier reveal that the accuracy for effectively detecting relevant eWOM for SentiStrength are rather inaccurate: 0.22 (Dissatisfaction), 0.19 (Neutral) and 0.07 (Satisfaction). Again, the results improve when employing the machine learning techniques, especially for the Satisfaction category. As for Satisfaction, SVM achieved the best result in terms of accuracy (0.52), followed by LR (0.51) and PA (0.50). Moreover, the accuracy of the various algorithms in the Neutral category ranges from 0.15 to 0.37. Overall, the best classification method was LR (0.37), followed by SVM (0.36) and PA (0.34). Finally, regarding the Dissatisfaction category, the SGD indicates the highest accuracy (0.39), followed by LR (0.35) and PA (0.25). In other words, the SGD is most effective when detecting urgent cases, namely consumers' complaints or rejections, on Facebook.[5]

4.2.2. Precision

Firstly, the results reveal that the precision for effectively detecting relevant eWOM when employing sentiment analysis for SentiStrength are: 0.07 (Satisfaction), 0.16 (Neutral) and 0.14 (Dissatisfaction). Precision is particularly high for the dictionary-based approach: 0.30 (Satisfaction), 0.35 (Neutral) and 0.41 (Dissatisfaction). The precision of the various machine learning techniques in the Satisfaction category varies between 0.38 and 0.67. The MNB indicates the best performance (0.67), followed by BNB (0.44) and

[5] By using the Akaike Information Criterion (AIC; the relative quality of a model for a given set of data), we estimated the quality of each algorithm for every eWOM type relative to each of the other algorithms. By doing so, we found decisive evidence in favour of MNB (Satisfaction), PA (Neutral) and BNB (Dissatisfaction) models relative to the other models (see Web Appendix F; Table F2 for more details).

Table 3
Model generation results of the generic classifier (eWOM type).

Category	Technique	Accuracy	Precision	Recall
Satisfaction ($N = 854$)				
Sentiment analysis	LIWC	0.05	0.06	0.04
	P	0.04	0.04	0.04
	SN	0.07	0.07	0.08
Dictionary-based	D	0.15	0.30	0.10
Machine learning	BNB	0.38	0.44	0.34
	MNB	0.32	0.67	0.21
	LR	0.51	0.38	0.76
	SGD	0.49	0.38	0.69
	SVM	0.52	0.41	0.63
	PA	0.50	0.40	0.68
Neutral ($N = 760$)				
Sentiment analysis	LIWC	0.13	0.16	0.10
	P	0.13	0.13	0.14
	SN	0.19	0.16	0.22
Dictionary-based	D	0.14	0.35	0.09
Machine learning	BNB	0.28	0.25	0.32
	MNB	0.15	0.34	0.10
	LR	0.37	0.25	0.74
	SGD	0.33	0.23	0.60
	SVM	0.36	0.24	0.69
	PA	0.34	0.24	0.60
Dissatisfaction ($N = 267$)				
Sentiment analysis	LIWC	0.20	0.15	0.29
	P	0.19	0.12	0.40
	SN	0.22	0.14	0.54
Dictionary-based	D	0.09	0.41	0.05
Machine learning	BNB	0.26	0.20	0.40
	MNB	0.25	0.48	0.16
	LR	0.35	0.23	0.77
	SGD	0.39	0.32	0.48
	SVM	0.04	0.02	1.00
	PA	0.35	0.23	0.71

Note. LIWC *Linguistic Inquiry and Word Count;* P *Pattern;* SN *Sentiment Net*; D *Dictionary-based;* BN *Bernoulli Naïve Bayes*; MNB *Multinomial Naïve Bayes*; LR *Logistic Regression*; SGD *Stochastic Gradient Descent*; SVM *Support Vector Machine*; and PA *Passive Aggressive*. Performance scores ≥0.60 have been highlighted. Results merely derived from the test set.

SVM (0.41). Additionally, the precision of the various algorithms in the Neutral category varies between 0.23 (SGD) and 0.34 (MNB). As for the Dissatisfaction category, the best classification method was MNB (0.48), followed by SGD (0.32), LR and PA (0.23).

4.2.3. Recall

The results for the sentiment analysis reveal that the recall for effectively detecting relevant eWOM SentiStrength are: 0.54 (Dissatisfaction), 0.22 (Neutral) and 0.08 (Satisfaction). The recall of the various machine learning algorithms in the Satisfaction category ranges from 0.21 to 0.76. The best classification method was LR (0.76), followed by SGD (0.69) and PA (0.68). When examining the Neutral category, the LR indicates the highest recall (0.74), followed by SVM (0.69), SGD and PA (0.60). As for Dissatisfaction, SVM (1.00), LR (0.77) and PA (0.71) indicate a better recall when compared to sentiment analysis. Interestingly, the SVM indicates a very low precision (0.02), while the recall indicates perfect learning (1.00). Thus, it returns many results, but most of its

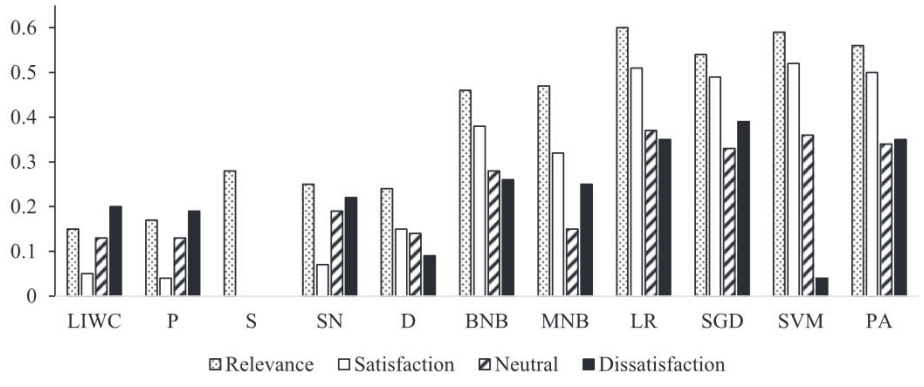

Fig. 2. Accuracy of Relevance and eWOM type (Generic Classifier).

predictions are incorrect when compared to the training set. Overall, the LR, SVM and PA perform better on recall, whereas the MNB and SGD better perform on precision (0.48 and 0.32, respectively).

Finally, Fig. 2 shows the accuracy of the generic classifier in terms of relevance and eWOM type per technique.

4.3. Industry-specific classifier for detecting relevant eWOM

Furthermore, we explored whether an industry-specific classifier is more accurate compared to a generic classifier in detecting relevant eWOM. We examined whether differences exist between various industries in detecting eWOM–and even detecting urgent eWOM–that requires a webcare response. Table 4 summarizes the results of the industry-specific classifier in terms of relevance. Since merely a few instances were found in the dataset for the Insurance industry, the results for an industry-specific classifier for the Insurance industry were unreliable and excluded from Table 4.

4.3.1. Accuracy

The accuracy of sentiment analysis per industry ranged from 0.10 to 0.46. For a number of industries, the results gradually improve when employing machine learning algorithms. In particular, LR achieved the best result in terms of accuracy for 'Food, Beverages & Tobacco', 'Telecommunication Services', 'Consumer Services' and 'Household & Personal Products' industries (Accuracy = 0.70, 0.68, 0.60 and 0.53, respectively).[6] Furthermore, the SVM achieved the best result in terms of accuracy for the 'Retailing' industry (0.64). Finally, the SVM and LR indicated the same accuracy for the industry of 'Automobiles & Components' (0.61). For all classifications, identifying relevant eWOM is the most effective for the 'Food, Beverages & Tobacco' (0.65), followed by 'Consumer Durables & Apparel' (0.60) and 'Automobiles & Components' (0.59). Overall, an industry-specific classifier is more effective than a generic classifier for the following industries: 'Food, Beverages & Tobacco', 'Consumer Durables & Apparel', 'Consumer Services', 'Automobiles & Components' and 'Telecommunication Services'.

4.3.2. Precision

The results indicate that the LR achieved the best precision for the 'Automobiles & Components' (0.60) and the SVM for 'Telecommunication Services' (0.62).

Furthermore, the PA achieved the best results in terms of precision for 'Consumer Durables & Apparel' (0.60), 'Food, Beverages & Tobacco' (0.65). Overall, the MNB achieved the best precision, particularly for the following industries: 'Consumer Services' (0.54), 'Retailing' (0.58) and 'Household & Personal Products' (0.59). As for the industries 'Food, Beverages & Tobacco' and 'Telecommunication Services', an industry-specific classifier indicates a comparable or higher precision than a generic classifier.

4.3.3. Recall

Furthermore, the results revealed that LR, SGD, SVM and PA indicate good learning in terms of recall. The LR achieved the best result for 'Consumer Services' (0.81), 'Retailing' (0.79), 'Telecommunication Services' (0.79), and 'Household & Personal Products' (0.78). Furthermore, the SVM achieved the best result in terms of recall for 'Consumer Durables & Apparel' (0.84). Finally, when examining the recall of the 'Food, Beverages & Tobacco' and 'Automobiles & Components' industries, LR and SVM retrieved a comparable recall (0.77 and 0.67, respectively). Overall, the recall of the industries 'Consumer Durables & Apparel', 'Retailing', 'Consumer Services' and 'Telecommunication Services' is higher when compared to a generic classifier.

4.4. Detecting relevant eWOM on another social media platform: Twitter

Finally, we examined whether the classifiers built with Facebook posts as training data are generalizable to other social media platforms, in this case Twitter. We used our subsample of tweets as a test set for our classifiers. Table 5 summarizes the results of a generic classifier via the different methods.

4.4.1. Accuracy

The results for Twitter, in terms of accuracy, are somewhat comparable to the results for Facebook.[7] The accuracy of the various algorithms ranged from 0.44 to 0.56 (Facebook: 0.46 to 0.60). The best classification method was SVM (0.56), followed by BNB (0.55), LR (0.55) and PA (0.55).

4.4.2. Precision

The precision of the various machine learning techniques varies between 0.32 and 0.53, which is somewhat lower than Facebook (i.e., 0.47 to 0.61). The MNB and PA indicate the best performance (0.53).

[6] By using the Akaike Information Criterion (AIC; the relative quality of a model for a given set of data), we estimated the quality of each algorithm for every industry relative to each of the other algorithms. By doing so, we found decisive evidence in favour of the LR model relative to the other models (see Web Appendix F; Table F3 for more details).

[7] By using the Akaike Information Criterion (AIC; the relative quality of a model for a given set of data), we estimated the quality of each algorithm relative to each of the other algorithms. By doing so, we found decisive evidence in favour of the SVM model relative to the other models (see Web Appendix F; Table F4 for more details).

Table 4
Model generation results of the industry-specific classifier (Relevance).

Industry	Technique	Accuracy	Precision	Recall
Automobiles & Components ($N = 178$)				
Sentiment Analysis	LIWC	.12	.44	.07
	P	.28	.20	.46
	S	.25	.17	.46
	SN	.25	.18	.42
Dictionary-based	D	.29	.70	.18
Machine Learning	BNB	.51	.55	.48
	MNB	.60	.59	.61
	LR	.63	.60	.67
	SGD	.57	.57	.57
	SVM	.63	.59	.67
	PA	.58	.58	.59
Consumer Durables & Apparel ($N = 86$)				
Sentiment Analysis	LIWC	.14	.44	.08
	P	.41	.33	.53
	S	.39	.28	.68
	SN	.46	.35	.68
Dictionary-based	D	.28	.42	.21
Machine Learning	BNB	.58	.58	.58
	MNB	.59	.55	.63
	LR	.62	.50	.80
	SGD	.60	.54	.67
	SVM	.65	.54	.84
	PA	.61	.60	.62
Consumer Services ($N = 402$)				
Sentiment Analysis	LIWC	.10	.28	.06
	P	.11	.07	.22
	S	.22	.13	.62
	SN	.22	.14	.49
Dictionary-based	D	.26	.29	.24
Machine Learning	BNB	.46	.48	.44
	MNB	.44	.54	.38
	LR	.61	.50	.81
	SGD	.55	.46	.68
	SVM	.60	.49	.79
	PA	.56	.47	.70
Food, Beverages & Tobacco ($N = 271$)				
Sentiment Analysis	LIWC	.11	.47	.06
	P	.33	.23	.58
	S	.29	.19	.62
	SN	.30	.21	.54
Dictionary-based	D	.18	.43	.11
Machine Learning	BNB	.56	.61	.52
	MNB	.63	.64	.62
	SGD	.63	.60	.66
	SVM	.68	.60	.77
	PA	.64	.65	.62
Household & Personal Products ($N = 402$)				
Sentiment Analysis	LIWC	.15	.28	.10
	P	.13	.08	.47
	S	.11	.06	.61
	SN	.12	.07	.53
Dictionary-based	D	.22	.29	.18
Machine Learning	BNB	.44	.42	.46
	MNB	.41	.59	.32
	LR	.52	.39	.78
	SGD	.45	.35	.64
	SVM	.52	.40	.73
	PA	.50	.41	.64
Retailing ($N = 150$)				
Sentiment Analysis	LIWC	.20	.37	.13
	P	.13	.08	.39
	S	.16	.09	.72
	SN	.14	.08	.50
Dictionary-based	D	.25	.28	.23
Machine Learning	BNB	.33	.35	.31
	MNB	.37	.58	.27
	LR	.51	.38	.79
	SGD	.54	.50	.58
	SVM	.53	.44	.67
	PA	.51	.51	.52

Table 4 (*continued*)

Industry	Technique	Accuracy	Precision	Recall
Telecommunication Services (*N* = 233)				
Sentiment Analysis	LIWC	.23	.40	.16
	P	.36	.29	.47
	S	.38	.26	.65
	SN	.35	.26	.53
Dictionary-based	D	.34	.38	.30
Machine Learning	BNB	.55	.59	.52
	MNB	.57	.58	.56
	LR	.69	.61	.79
	SGD	.60	.57	.62
	SVM	.69	.62	.77
	PA	.63	.61	.64

Note. LIWC *Linguistic Inquiry and Word Count;* P *Pattern;* S¸ *Sentiment Negative*; SN¸ *Sentiment Net*; D *Dictionary-based;* BNB *Bernoulli Naïve Bayes*; MNB *Multinomial Naïve Bayes*; LR *Logistic Regression*; SGD *Stochastic Gradient Descent*; SVM *Support Vector Machine*; and PA *Passive Aggressive*. Performance scores ≥0.60 have been highlighted. Results are merely derived from the test set.

4.4.3. Recall

Finally, the recall of the various machine learning techniques varies between 0.38 and 0.80, which is comparable to Facebook (i.e., 0.38 and 0.79). The best classification method was LR (0.80), followed by SGD (0.76) and BNB (0.72). Overall, the results indicate that the classifiers that have been trained on Facebook data, are generating somewhat comparable results on Twitter. As tweets are (usually) shorter, sentiment analysis is more effective in detecting eWOM in need of a webcare response compared to Facebook, but the classifiers compute comparable results between the two social media platforms and still outperform any type of sentiment analysis and dictionary-based approaches.

5. General discussion

5.1. Theoretical implications

This study sets out to explore how machine learning can assist firms and brands with reliably identifying relevant eWOM that would require a response. Using a sample of over 60,000 Facebook posts and approximately 11,000 tweets, for 16 brands across eight industry segments, we compared the effectiveness of traditional methods of detecting posts requiring a response–i.e., four types of sentiment analysis and a dictionary-based approach–with the usage of a variety of machine learning algorithms, trained specifically for this purpose. These comparisons provide several important theoretical and managerial implications.

First and foremost, this study is a pioneering attempt to automatically detect relevant eWOM messages that require a webcare response in two different social media platforms. eWOM messages are characterized by unstructured text formats, text speak and simplified expressions. Social media monitoring tools based on sentiment analyses have faced important gaps when analysing eWOM content (Pai et al., 2013). As many social media monitoring tools rely on extracting sentiment out of text, they are incapable of dealing with constructs in which the literal meaning of the text is not the intended meaning of the eWOM content

Table 5
Model generation results of the generic classifier on Twitter data (Relevance).

Technique (*N* = 11,154)	Accuracy	Precision	Recall
Sentiment analysis			
LIWC	0.25	0.51	0.16
P	0.30	0.38	0.25
S	0.26	0.43	0.19
SN	0.12	0.47	0.07
Dictionary-based			
D	0.34	0.53	0.25
Machine learning			
BNB	0.55	0.45	0.72
MNB	0.44	0.53	0.38
LR	0.55	0.41	0.80
SGD	0.45	0.32	0.76
SVM	0.56	0.46	0.71
PA	0.55	0.53	0.56

Note. LIWC *Linguistic Inquiry and Word Count;* P *Pattern;* S¸ *Sentiment Negative*; SN¸ *Sentiment Net*; D *Dictionary-based;* BNB *Bernoulli Naïve Bayes*; MNB *Multinomial Naïve Bayes*; LR *Logistic Regression*; SGD *Stochastic Gradient Descent*; SVM *Support Vector Machine*; and PA *Passive Aggressive*. Performance scores ≥0.60 have been highlighted.

(Kunneman et al., 2015). Our findings indicate that supervised machine learning can be an effective approach, achieving levels of precision and recall twice as high as the levels achieved by sentiment analysis when detecting whether an eWOM message requires a response or not. It is noteworthy that our approach also outperforms sentiment analysis techniques that are based on machine learning. This further underscores the conceptual inexpedience of sentiment extraction as a means for monitoring eWOM in a webcare context.

Second, this study revealed that models using machine learning can detect and categorize relevant eWOM more effectively compared to sentiment analysis. Our machine learning based categorization showed that supervised machine learning models trained specifically for eWOM detection and categorization outperform sentiment analysis both when it comes to identifying response-worthy eWOM, and when it comes to categorizing eWOM in satisfaction, dissatisfaction or neutral categories. Moreover, while we mainly benchmarked our dissatisfaction categories against negative eWOM categories from different sentiment analysis and dictionary-based approaches, our approach acknowledges the fact that – conceptually – a sole focus on negative sentiment is not sufficient in a webcare context (e.g., Schamari & Schaefers, 2015). Therefore, beyond being superior to all types of sentiment analyses in identifying dissatisfaction, our approach also identifies relevant eWOM that is positive or neutral but would yet require a response. Importantly, our study showed that 84.2% of all relevant eWOM on Facebook was either neutral or positive (58.8% for Twitter). Not being able to indicate the relevance of this type of eWOM is a major shortcoming of any kind of sentiment analysis as brands risk to alienate consumers if they do not engage in dialogue with them (Colliander et al., 2015).

Third, industry-specific classifiers are more accurate compared to a generic classifier in detecting relevant eWOM. Particularly, when examining the effectiveness of machine learning techniques, industry-specific classifiers for 'Consumer Durables & Apparel', 'Food, Beverages & Tobacco', 'Telecommunication Services', 'Insurance', 'Consumer Services' and 'Automobiles & Components' have indicated an increase in the effectiveness of relevant eWOM detection, while generic classifiers better classified the remaining industries. This further supports the notion that not all eWOM is created equally (e.g., Marchand, Hennig-Thurau, & Wiertz, 2017) and that brands should seek for monitoring solutions that fit their own purpose, instead of relying on off-the-shelve solutions.

5.2. Managerial implications

The current study also has important managerial implications. First of all, the findings clearly demonstrate that supervised machine learning can be a more effective approach than sentiment analysis for detecting whether an eWOM message requires a response or not. Also, this study revealed that models using machine learning can detect and categorize relevant eWOM more effectively compared to sentiment analysis. With machine learning consumer feedback can be categorized in a more useful way, instead of just having an indication of the sentiment, consumer feedback can be categorized into three customer satisfaction dimensions. For a firm this means that, by adopting machine learning techniques, different types of consumer feedback can be addressed in organizations by different departments or teams that actually have the expertise to deal with the specific type of feedback. For example, rejections and complaints might be handled by customer care, whereas suggestions could be handled by R&D, while compliments might be most suited to be taken care of by marketing staff. This approach might be much more efficient than having all types of consumer feedback handled by one webcare team.

Second, the findings also demonstrate that industry-specific classifiers were more accurate than the generic classifiers. However, managerial considerations must first be taken into account in order to decide which machine learning algorithm will have the best performance. Precision and recall refer to how useful and complete the search results are, respectively. The MNB has shown to be more effective compared to other classifiers. Though, when it is crucial for a brand to identify all relevant eWOM a higher recall is desirable. Then, the LR, SGD, SVM and PA are suggested. Approximately 85% of Facebook content and 70% of Twitter content in the sample was not in need of a webcare response. Supervised machine learning has shown its ability to help marketers with detecting relevant eWOM in the cluttered social media landscape (Grégoire et al., 2015).

Third, as industry specific classifiers were more accurate than the generic classifiers, classifiers built for specific product categories might be more accurate as well. Therefore, for house of brands businesses, such as Unilever and Procter & Gamble, it might be important to build classifiers for different product or brand categories and have consumer feedback decentrally organized.

Fourth, the findings demonstrated that, when on Facebook, an overwhelming amount of eWOM messages has been written by satisfied consumers (44%), while on Twitter around 40% was written by dissatisfied consumers. Although complaint handling is of essential importance for webcare teams, in terms of customer relationship and reputation management, detecting eWOM messages of dissatisfied consumers with machine learning proved to be more difficult as compared to detecting eWOM messages of satisfied consumers. In light of this, one of the realistic strategies for firms is to be more selective with negative eWOM (Van Noort & Willemsen, 2012) and more complaisant to positive eWOM (e.g., Demmers, Van Dolen, & Weltevreden, 2018; Schamari & Schaefers, 2015).

Furthermore, based on the current findings, brands are strongly encouraged to use machine learning techniques instead of sentiment analysis-based approaches. On the one hand, brands could use the power of eWOM when consumers are energetic endorsers of positive feedback. The fact that satisfied consumers are willing to share their positive experiences means that their loyalty to the brand might be greater after they comment (Okazaki et al., 2015). This could provide brands with a benefit in market competitiveness and brand loyalty. On the other hand, dissatisfied consumers who share their opinion on social media tend to be upset and disappointed about their experience. These types of opinions may seriously harm a brand's reputation and thus need to be taken care of as soon as possible. Firms should not only carefully monitor, but also examine and analyse what is being said

about their brands on social network sites so that satisfied consumers' ability to produce goodwill will not be disturbed, confused, or damaged by dissatisfied consumers.

Social media monitoring based on machine learning is relatively easy to implement. In our case, coders spent approximately 300 h to code about 60,000 Facebook comments and 11,000 tweets. Since the tweets were only coded to test the already trained algorithm, we calculate a net coding time of 250 h to implement a technique that we prove to be broadly applicable. In many cases brands might also use their own databases, if existing, which may already include (some of) the necessary categorizations (for example, an overview of posts on the brand's Facebook page, and which posts their webcare teams decided to respond to). The benefits are great, considering increases in precision and recall of up to 100%. Moreover, we assume our approach to be highly efficient in saving time for webcare staff, because a large portion of irrelevant eWOM will be filtered out (up to 85% in this study) before showing up in a webcare dashboard. Overall, the cost-benefit ratio of our approach appears to be very promising.

5.3. Limitations and future research

In order to interpret our results, a few limitations should be recognized. First, in our method we identified generic types of eWOM, rather than specific issues which might be more related to specific products or brands, to increase the generalizability of the results to other platforms. We found variance in the generic distinction between Dissatisfaction (i.e., consumers expressing a rejection and/or a complaint), Satisfaction (i.e., acknowledgements and/or compliments) and Neutral eWOM (comments, questions and/or suggestions). Also, for Facebook we demonstrated that an industry specific classifier was more accurate than a generic classifier. In a similar vein, classifiers built for specific brand-related or product-related topics in eWOM, might be more accurate. Therefore, we may need to refine the generic classification to capture more detailed communication patterns in eWOM or to identify specific topics that are relevant for specific brands or products.

Second, this study merely examined eWOM messages on brand pages on Facebook and on Twitter. Future research should examine the possibilities of automatically detecting webcare on other platforms, such as review sites, news websites, etc. as a wide range of UGC websites are being used for the online opinion exchange between consumers. The current data clearly demonstrates that between Facebook and Twitter, percentages of eWOM type differ greatly: the amount of satisfied eWOM was a lot higher for Facebook, while the amount of dissatisfied eWOM was higher for Twitter. Furthermore, on Twitter relatively more questions (Neutral eWOM) were asked. Future research could investigate how eWOM type differs between a multitude of platforms, to gain further insights into how platforms are used for opinion sharing. Such research could also further inform managers: Machine learning can be of great help on platforms demonstrating variance in eWOM type—i.e., the extent to which consumer posts require a response or are satisfied, neutral and dissatisfied messages. If almost all messages express require a response or express dissatisfaction—for example in a brand owned channel explicitly created to handle complaints or webcare—machine learning models would be better used to directly categorize consumer feedback into more specific types of messages (e.g., complaints, question, and so on) instead of whether it requires a response in the first place.

Third, we chose to ignore comments that did not include an expression concerning the product, service and/or the entire brand, as these were considered irrelevant for a webcare response, and thus non-relevant for our study purpose. While for webcare-relevant comments consumers' motives are for example altruism, dissatisfaction, and restoration of justice, consumers' motives for webcare-irrelevant comments might be very different (e.g., self-expression and image-building). It might be interesting to explore motivations behind webcare-irrelevant comments as the amount of these comments was considerable, in both social media platforms and especially within Facebook.

6. Conclusion

This study seeks to incorporate various automated content analysis techniques into the domain of Facebook and Twitter. Applying these techniques, the relevance and eWOM type of social media messages have been classified. Our intent to examine whether machine learning techniques are more effective than sentiment analyses and dictionary-based approaches in detecting relevant eWOM was successful: we can conclude that LR, SGD and PA classifiers, in particular, are most accurate. Principally, this study revealed that detecting relevant eWOM is especially effective for satisfied consumers. Nonetheless, it is more difficult to detect eWOM messages of dissatisfied consumers. Further, this study has shown that for specific industries domain specificity could produce more accurate results compared to a generic classifier for webcare detection. Finally, we found that the classifiers we built on Facebook data were able to compute comparable results on Twitter.

The results do not only provide important insights regarding if and how consumer posts requiring a webcare response can be best automatically identified, but also allow for an exploration across a varied sample of industry sectors and platforms. Our findings may serve as an interesting stepping-stone for future research on automatically analysing eWOM.

Acknowledgements

The data in this study has been acquired in collaboration with Dentsu Aegis Network. The authors would like to thank Dentsu Aegis Network for this pleasant collaboration. This work was carried out on the Dutch National e-infrastructure with the support of SURF Cooperative (HPC Cloud).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2019.01.010.

References

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, *53*(3), 297–318.

Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, *29*(5), 815–827.

Bhandari, M., & Rodgers, S. (2018). What does the brand say? Effects of brand feedback to negative eWOM on brand trust and purchase intentions. *International Journal of Advertising*, *37*(1), 125–141.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge, UK: Cambridge University Press.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953–975.

Colicev, A., Malshe, A., Pauwels, K., & O'Connor, P. (2018). Improving consumer mindset metrics and shareholder value through social media: The different roles of owned and earned media. *Journal of Marketing*, *82*(1), 37–56.

Colliander, J., Dahlén, M., & Modig, E. (2015). Twitter for two: Investigating the effects of dialogue with customers in social media. *International Journal of Advertising*, *34*(2), 181–194.

Coombs, W. T. (2002). Assessing online issue threats: Issue contagions and their effect on issue prioritisation. *Journal of Public Affairs*, *2*(4), 215–229.

Coosto (2017). Coosto. Retrieved from https://www.coosto.com/nl.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cui, H., Mittal, V., & Datar, M. (2006, July). Comparative experiments on sentiment classification for online product reviews. *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1265–1270).

Demmers, J., Van Dolen, W. M., & Weltevreden, J. W. (2018). Handling consumer messages on social networking sites: Customer service or privacy infringement? *International Journal of Electronic Commerce*, *22*(1), 8–35.

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, *3*, 1265–1287.

Esuli, A., & Sebastiani, F. (2010). Machines that learn how to code open-ended survey data. *International Journal of Market Research*, *52*(6), 775–800.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, *23*(10), 1498–1512.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*(4), 545–560.

Gopinath, S., Thomas, J. S., & Krishnamurthi, L. (2014). Investigating the relationship between the content of online word of mouth, advertising, and brand performance. *Marketing Science*, *33*(2), 241–258.

Grégoire, Y., Salle, A., & Tripp, T. M. (2015). Managing social media crises with your customers: The good, the bad, and the ugly. *Business Horizons*, *58*(2), 173–182.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism and Mass Communication Quarterly*, *93*(2), 332–359.

Hansen, N., Kupfer, A. K., & Hennig-Thurau, T. (2018). Brand crises in the digital age: The short-and long-term effects of social media firestorms on consumers and brands. *International Journal of Research in Marketing*, *35*(4), 557–574.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2018). Comparing automated text classification methods. *International Journal of Research in Marketing*, 1–19 (Advance online publication).

Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, *52*(5), 629–641.

Huibers, J., & Verhoeven, J. (2014). Webcare als online reputation management. *Tijdschrift voor Communicatiewetenschap*, *42*(2), 165–189.

Humphreys, A., & Wang, R. J. H. (2017). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274–1306.

Kaiser, C., & Bodendorf, F. (2012). Mining consumer dialog in online forums. *Internet Research*, *22*(3), 275–297.

Kent, A., Berry, M. M., Luehrs, F. U., Jr., & Perry, W. (1955). Operational criteria for designing information retrieval systems. *American Documentation*, *6*(2), 93–108.

Kietzmann, J., & Canhoto, A. (2013). Bittersweet! Understanding and managing electronic word of mouth. *Journal of Public Affairs*, *13*(2), 146–159.

Kübler, R. V., Colicev, A., & Pauwels, K. H. (2017). Social media's impact on consumer mindset: When to use which sentiment extraction tool. *Marketing Science Institute Working Paper Series*, *17*(122), 1–99.

Kübler, R. V., Wieringa, J. E., & Pauwels, K. H. (2017). Machine learning and big data. In Leeflang, Wieringa, Bijmolt, & Pauwels (Eds.), *Advanced methods for modeling markets* (pp. 631–670). Berlin: Springer.

Kunneman, F., Liebrecht, C., Van Mulken, M., & Van Den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management*, *51*(4), 500–509.

Kupor, D., & Tormala, Z. (2018). When moderation fosters persuasion: The persuasive power of deviatory reviews. *Journal of Consumer Research*, *45*(3), 490–510.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, *48*(2), 354–368.

Loper, E., & Bird, S. (2002, July). NLTK: The natural language toolkit. *Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics* (pp. 63–70).

Marchand, A., Hennig-Thurau, T., & Wiertz, C. (2017). Not all digital word of mouth is created equal: Understanding the respective impact of consumer reviews and microblogs on new product success. *International Journal of Research in Marketing*, *34*(2), 336–354.

Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, *20*(1), 1–28.

Maslowska, E., Malthouse, E. C., & Bernritter, S. F. (2017). Too good to be true: The role of online reviews' features in probability to buy. *International Journal of Advertising*, *36*(1), 142–163.

Moon, S., & Kamakura, W. A. (2017). A picture is worth a thousand words: Translating product reviews into a product positioning map. *International Journal of Research in Marketing*, *34*(1), 265–285.

Munoz-Garcia, O., & Navarro, C. (2012). Comparing user-generated content published in different social media sources. *NLP can u tag# user generated-content conference on language resources and evaluation (LREC)* (pp. 1–8) Retrieved from: https://www.slideshare.net/omunozgarcia/comparing-user-generated-content-published-in-different-social-media-sources.

Nakhaeizadeh, G., Taylor, C. C., & Kunisch, G. (1997). Dynamic supervised learning: Some basic issues and application aspects. *Classification and knowledge organization* (pp. 123–135). Berlin, Heidelberg: Springer.

Okazaki, S., Diaz-Martin, A. M., Rozano, M., & Menendez-Benito, H. D. (2015). How to mine brand tweets: Procedural guidelines and pre-test. *International Journal of Market Research*, *56*(4), 467–488.

Ordenes, F. V., Grewal, D., Ludwig, S., Ruyter, K. D., Mahr, D., & Wetzels, M. (2018). Cutting through content clutter: How speech and image acts drive consumer sharing of social media brand messages. *Journal of Consumer Research*, 1–25 forthcoming.

Owsley, S., Sood, S., & Hammond, K. J. (2006). Domain specific affective classification of documents. *Proceedings of the AAAI Symposium on computational approaches to analysing weblogs* (pp. 181–183).

Packard, G., & Berger, J. (2017). How language shapes word of mouth's impact. *Journal of Marketing Research*, *54*(4), 572–588.

Pai, M., Chu, H., Wang, S., & Chen, Y. (2013). Electronic word of mouth analysis for service experience. *Expert Systems with Applications*, *40*, 1993–2006.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.
Pauwels, K., Aksehirli, Z., & Lackman, A. (2016). Like the ad or the brand? Marketing stimulates different electronic word-of-mouth content to drive online and offline performance. *International Journal of Research in Marketing*, *33*(3), 639–655.
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* Austin, TX: University of Texas at Austin.
Phillips, R. L., & Ormsby, R. (2016). Industry classification schemes: An analysis and review. *Journal of Business and Finance Librarianship*, *21*(1), 1–25.
Schamari, J., & Schaefers, T. (2015). Leaving the home turf: How brands can use webcare on consumer-generated platforms to increase positive consumer engagement. *Journal of Interactive Marketing*, *30*, 20–33.
Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, *51*(4), 387–402.
Tang, T., Fang, E., & Wang, F. (2014). Is neutral really neutral? The effects of neutral user- generated content on product sales. *Journal of Marketing*, *78*(4), 41–58.
Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198–215.
Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463–479.
Van Laer, T., & De Ruyter, K. (2010). In stories we trust: How narrative apologies provide cover for competitive vulnerability after integrity-violating blog posts. *International Journal of Research in Marketing*, *27*, 164–174.
Van Noort, G., & Willemsen, L. M. (2012). Online damage control: The effects of proactive versus reactive webcare interventions in consumer-generated and brand-generated platforms. *Journal of Interactive Marketing*, *26*(3), 131–140.
Van Noort, G., Willemsen, L. M., Kerkhof, P., & Verhoeven, J. W. M. (2014). Webcare as an integrative tool for customer care, reputation management, and online marketing: A literature review. In P. J. Kitchen, & E. Uzunoğlu (Eds.), *Integrated communications in the postmodern era* (pp. 77–99). Basingstoke, Hampshire: Palgrave Macmillan.
Weitzl, W., Hutzinger, C., & Einwiller, S. (2018). An empirical study on how webcare mitigates complainants' failure attributions and negative word-of-mouth. *Computers in Human Behavior*, *89*, 316–327.
Willemsen, L., Neijens, P. C., & Bronner, F. A. (2013). Webcare as customer relationship and reputation management? Motives for negative electronic word of mouth and their effect on webcare receptiveness. *Advances in Advertising Research*, *4*, 55–69.
Willemsen, L. M., Neijens, P. C., Bronner, F., & De Ridder, J. A. (2011). "Highly recommended!" The content characteristics and perceived usefulness of online consumer reviews. *Journal of Computer-Mediated Communication*, *17*(1), 19–38.
Wilson, A. E., Giebelhausen, M. D., & Brady, M. K. (2017). Negative word of mouth can be a positive for consumers connected to the brand. *Journal of the Academy of Marketing Science*, *45*(4), 534–547.
Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi- words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765.
Zhang, Z., Li, X., & Chen, Y. (2012). Deciphering word-of-mouth in social media: Text- based metrics of consumer reviews. *ACM Transactions on Management Information Systems (TMIS)*, *3*(1), 1–23.